6Г

УДК 004.274:004.056

С.Я. Гильгурт
Институт проблем моделирования в энегргетике им. Г.Е. Пухова НАН Украины, г. Киев
Украина, 03164, г. Киев, ул. Генерала, Наумова, 15, *hilgurt@ukr.net*

Аппаратное распознавание строк в интеллектуальных системах защиты информации

S.Ya. Hilgurt
Pukhov Institute for Modelling in Energy Engineering, NAS of Ukraine
Ukraine, 03164, Kiev, General Naumov st., 15

Hardware String Matching in Intellectual Security Systems

С.Я. Гільгурт
Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України, м. Київ
Україна, 03164, м. Київ, вул. Генерала Наумова, 15

Апаратне розпізнавання рядків в інтелектуальних системах захисту інформації

При создании интеллектуальных систем противодействия таким угрозам информационной безопасности, как сетевые вторжения, вирусы и спам, необходимо анализировать интенсивный поток данных на наличие одновременно нескольких тысяч эталонных последовательностей символов. Для достижения требуемой производительности часто используют аппаратные решения на базе программируемых интегральных схем. В настоящей работе исследован зарубежный опыт подобных разработок, предложено применение унифицированных изделий.
Ключевые слова: множественное распознавание строк, информационная безопасность, ПЛИС, реконфигурируемый вычислитель

In order to protect information systems from security threats such as intrusion, virus and spam it is necessary to match all occurrences of a predefined set of string-based patterns containing several thousands of strings. To provide required throughput, the hardware solutions based on programmable logic are widely used. In this paper, the world experiences of such works are investigated and unified solution is proposed.
Key words: multi-pattern string matching, information security, FPGA, reconfigurable accelerator

При створенні інтелектуальних систем протидії таким загрозам інформаційній безпеці, як мережні вторгнення, віруси та спам, необхідно аналізувати інтенсивний потік даних на наявність одночасно декількох тисяч еталонних послідовностей символів. Для досягнення необхідної продуктивності часто використовують апаратні рішення на базі програмованих інтегральних схем. У даній роботі досліджений зарубіжний досвід подібних розробок, запропоновано застосування уніфікованих виробів.
Ключові слова: множинне розпізнавання рядків, інформаційна безпека, ПЛІС, реконфігурований обчислювач

Введение

Задача поиска заданной подстроки в потоке символов, как частный случай более общей задачи распознавания образов, возникает во многих прикладных областях, связанных с обработкой информации, таких, как интеллектуальный анализ данных (data mining), оптимизация транзакций в СУБД, измерение скоростных характеристик и оперативный

мониторинг сетевого трафика, поддержка технологий QoS (Quality of Service – качество обслуживания) и IP-телефонии, распознавание цепочек атомов и их комбинаций в молекулярной биологии и др.

Актуальна данная проблема и при создании систем информационной безопасности. Непрерывно растущие объемы данных и сетевого трафика в корпоративных, удаленных и распределенных сетях (грид, облачные вычисления) при остановившемся росте тактовой частоты процессорных устройств снижают эффективность программных решений. Как следствие, только аппаратные компоненты способны обеспечить приемлемые показатели производительности и время реакции для современных систем защиты данных.

Анализ информационных источников свидетельствует о наличии большого количества англоязычных разработок по аппаратным средствам защиты информации, в том числе на базе программируемых логических интегральных схем (ПЛИС). Однако в литературе, издаваемой в странах СНГ, фактически отсутствуют публикации на данную тему.

В настоящей работе проанализирован мировой опыт и особенности использования реконфигурируемых устройств для реализации аппаратного распознавания строк в системах защиты информации. Для повышения эффективности процесса разработки систем распознавания на базе программируемой логики предложено использование типовых изделий – реконфигурируемых унифицированных вычислителей.

Задачи распознавания в информационной безопасности

Необходимость поиска заданных образцов строк в интенсивном потоке данных возникает при решении различных задач информационной безопасности, таких как обнаружение вторжений, антивирусная защита, борьба со спамом. Актуальна она и для появившегося относительно недавно направления, связанного с предотвращением потери корпоративных данных, – так называемые DLP-системы (Data Loss Prevention).

Поскольку перечисленные задачи, как правило, приходится решать в комплексе, возникли и интенсивно развиваются интегрированные системы информационной безопасности – Unified Threat Management (UTM). Такие системы объединяют в себе функциональность межсетевых экранов, антивирусов, спам-фильтров, средств информационной защиты контента.

Несмотря на различное назначение функций, выполняемых UTM-системами, при их реализации приходится решать сходную задачу выявления заданной строки (точнее, набора строк) в информационном потоке.

Исторически первыми разработками в сфере защиты информации, для которых началось практическое освоение аппаратных подходов к распознаванию строк, явились системы обнаружения вторжений (СОВ). Как следствие, данная область оказалась наиболее исследованной, для нее имеется большое число успешных наработок [1]. По этой причине в настоящей работе решение задачи распознавания строк с применением реконфигурируемых устройств рассматривается на примере СОВ.

Системы обнаружения вторжений

Структура СОВ в зависимости от назначения и особенностей применения может состоять из различных компонентов [2], но в обязательном порядке содержит один или несколько сенсоров. В состав сенсора, в свою очередь, обязательно входит модуль обнаружения атак, который выполняет ресурсоемкую операцию распознавания строк в интенсивном потоке сетевых пакетов.

Алгоритм функционирования системы обнаружения атак в общем случае состоит из трех этапов:

– захват сетевых пакетов (packet capture);
– фильтрация и сборка пакетов (filtering / fragmentation reassembly);
– распознавание (pattern matching).

Самым ресурсоемким является последний этап, который сводится к выполнению большого количества операций сравнения содержимого сетевых пакетов со строками символов, содержащимися в базе данных признаков известных атак (сигнатур).

Анализ сетевого трафика может осуществляться двумя способами: путем тотального захвата и инспектирования всех пакетов либо с учетом сетевых протоколов (с разборкой заголовков пакетов). Системы, основанные на первом способе, распознают большее число атак, но они значительно сложнее в своей реализации. По этой причине в исследуемых источниках преимущественно описываются системы, учитывающие протоколы.

База данных известных атак помимо эталонных последовательностей строк содержит правила их распознавания. Но обработка правил (анализ заголовков сетевых пакетов) производится по алгоритмам, которые давно и успешно применяются в межсетевых экранах. В исследовательском плане намного больший интерес представляет собственно задача поиска строк в потоке данных (анализ содержимого сетевых пакетов).

Ключевым компонентом СОВ является модуль распознавания, выполняющий самую ресурсоемкую вычислительную операцию поиска. От качества его реализации в значительной степени зависят такие важные характеристики системы обнаружения вторжений, как производительность, ресурсоемкость и масштабируемость [3].

В связи с постоянным ростом числа и сложности компьютерных атак, а также из-за значительного увеличения объемов данных, передаваемых в сети, в качестве аппаратной основы СОВ получили широкое распространение ПЛИС типа Field Programmable Gate Array (FPGA) [4].

Современные СБИС программируемой логики, содержащие миллионы эквивалентных логических элементов на одном кристалле, позволяют достичь скорости обработки информационного потока в несколько Гбит/сек для баз эталонов емкостью в несколько тысяч записей [5], [6].

Рассмотрим структурные и функциональные особенности систем обнаружения вторжений, а также основные принципы применения программируемой логики для их построения.

Требования, предъявляемые к СОВ на базе ПЛИС

Проведенный анализ известных разработок позволяет сформулировать требования, предъявляемые к аппаратным средствам на базе ПЛИС, используемых в качестве платформы для построения систем обнаружения вторжений, а также основные параметры, по которым следует оценивать их эффективность [7-9].

Главными показателями производительности СОВ являются: *максимальное число строк*, распознаваемых системой, и *пропускная способность*, которая может при этом быть достигнута.

Однако на практике намного более важной и труднодостижимой характеристикой архитектуры СОВ оказалась *масштабируемость* – способность наращивать возможности в широких пределах без несоизмеримо высоких дополнительных затрат. Актуальность данного свойства технического решения системы обнаружения вторжений обусловлена, с одной стороны, стремительным ростом сетевого трафика, с другой – постоянным увеличением размеров базы данных сигнатур.

Важной характеристикой СОВ, также связанной с производительностью, является *предсказуемость пропускной способности*, то есть независимость ее временных характеристик от состава входных данных. Обнаружение злонамеренного контента в сетевом трафике является редким событием, вероятность возникновения которого в штатном режиме невелика. Однако если содержимое анализируемых сетевых пакетов существенно влияет на быстродействие модуля распознавания СОВ, то такая система может оказаться уязвимой к намеренному засорению злоумышленником сетевого трафика элементами сигнатур известных атак.

Специфической чертой систем обнаружения вторжений, принцип действия которых основан на распознавании сигнатур, является необходимость регулярного обновления активной базы данных. *Удобство динамического обновления* существенно влияет на практическую полезность технического решения. Данный показатель затрагивает такие моменты, как возможность обновления базы сигнатур без приостановки процесса распознавания, способность обходиться без перепрограммирования ПЛИС, либо, в противном случае, наличие средств автоматической генерации и загрузки в ПЛИС новой конфигурации, а также удобство и скорость выполнения данной операции.

Независимость от состава базы данных сигнатур также является важной характеристикой СОВ. Ориентация модуля распознавания на ограниченный алфавит с целью повышения быстродействия может привести к нежелательным последствиям при его использовании в распознающих системах.

Помимо скоростных характеристик систем обнаружения вторжений, для их практического использования важны также стоимостные показатели. *Объем оперативной памяти*, необходимой для реализации выбранного алгоритма распознавания существенно влияет, в итоге, на быстродействие. Если имеющихся в ПЛИС ресурсов быстродействующей блочной памяти (BRAM) недостаточно для реализации запоминающего устройства, то возникает необходимость во внешней памяти, которая намного медленнее внутренней.

Существенной является также *общая стоимость реализации* системы. Каким бы эффективным ни был модуль распознавания, если для его интеграции в СОВ необходимы существенные дополнительные затраты, например, на преобразование формы представления информации, общая стоимость решения может оказаться неудовлетворительной.

Параллельное распознавание строк

Как указывалось выше, модуль распознавания строк является наиболее важным компонентом СОВ, от успешной реализации которого во многом зависят рассмотренные показатели эффективности. Следовательно, выбор алгоритма распознавания и технического решения для его реализации являются ключевыми моментами при создании систем обнаружения вторжений.

Однако задаче множественного распознавания строк в значительной степени свойственен параллелизм, причем, по двум направлениям: во-первых, несколько сетевых пакетов могут анализироваться одновременно; во-вторых, сравнение может производиться сразу со многими подстроками из базы данных сигнатур [10]. Рассмотрим эти направления возможного распараллеливания.

К сожалению, при реализации параллелизма первого типа возникает трудноразрешимое противоречие: разделение интенсивного входного потока на большое число отдельных блоков, обрабатываемых независимыми вычислительными модулями, приводит к задержкам, пропорциональным коэффициенту распараллеливания, обусловлен-

ным большим размером блоков; уменьшение же размера блоков снижает полезность распараллеливания из-за эффекта перекрытия, тем более ощутимого, чем длиннее искомые подстроки. К тому же такой подход требует реализации сложных процессов управления, планирования и буферизации [9].

Распараллеливание по базе сигнатур, то есть разделение набора распознаваемых подстрок на подгруппы также возможно. Но в этом случае обнаруживается важная особенность – большое число сигнатур во многом повторяют друг друга.

Причем данный эффект самоподобия в силу конечности алфавита теоретически должен возрастать по мере роста баз данных сигнатур. Учет такого эффекта позволяет существенно повысить производительность распознающей подсистемы. Однако упомянутые выше одношаблонные алгоритмы непригодны для данной цели.

Таким образом, возникает теоретическая проблема множественного распознавания строк [9]. Ее суть заключается в одновременном поиске во входной последовательности символов не одной подстроки, а заданного набора подстрок, различные фрагменты которых повторяются в значительной степени. Известные способы распараллеливания не позволяют добиться приемлемого результата в силу указанных выше причин. Следовательно, эффективное решение данной задачи можно получить лишь на уровне алгоритма или вычислительной структуры.

Подходы к построению модуля распознавания

В существующих на сегодняшний день системах аппаратного распознавания строк задействованы разнообразные подходы, приемы и технические решения. Наиболее распространенными из них являются [9]:

– цифровые автоматы;
– параллельные дискретные компараторы;
– устройства ассоциативной памяти и ее разновидности;
– различные варианты использования хэш-функций, в частности, фильтры Блума.

Для большинства подходов возможно применение конвейеризации. При кластеризации словаря эталонных строк используют методы теории графов.

Каждое из упомянутых направлений имеет как некоторые преимущества перед другими, так и недостатки. Например, цифровые автоматы, синтезированные в ПЛИС, не обеспечивают высокую пропускную способность, сложны в построении и конфигурировании. Параллельные компараторы при большей производительности приводят к повышенным затратам оборудования и плохой масштабируемости. Решения, основанные на ассоциативной памяти, менее требовательны к ПЛИС, чем цифровые компараторы при соизмеримой производительности, но дороже и потребляют больше энергии. Фильтр Блума и сокращение аппаратных затрат функциями кэширования позволяют уменьшить число сравнений, но обеспечивают вероятностное распознавание, что требует дополнительных затрат на доуточнение результатов совпадения.

Таким образом, ни один из упомянутых направлений не удовлетворяет в полной мере сформулированным выше требованиям, предъявляемым к системам обнаружения вторжений.

Следует также отметить наметившуюся в последние годы тенденцию к объединению в одном устройстве нескольких подходов и решений. При этом наиболее эффективным оказывается такое комбинирование, при котором учитываются особенности конкретного словаря эталонов, в частности, производится ранжирование распознаваемых строк по их длине.

Реконфигурируемые унифицированные вычислители

По мере увеличения числа и сложности сетевых компьютерных атак, а также из-за прекратившегося роста частоты микропроцессоров, программная реализация систем обнаружения вторжений становится все более проблематичной. Размеры баз сигнатур современных СОВ исчисляются тысячами записей. Именно по этой причине в настоящее время существенно возрос интерес к реализации данных систем на базе программируемых интегральных схем типа FPGA [11]. Высокая гибкость программируемой логики в сочетании с быстродействием аппаратного решения позволяют эффективно использовать естественный параллелизм, присущий задаче распознавания строк, которая является наиболее ресурсоемкой операцией в современных системах обнаружения вторжений.

Одна из главных трудностей практического применения программируемой логики обусловлена высокой стоимостью и трудоемкостью процесса разработки реконфигурируемых ускорителей, присущая любому аппаратному решению. С другой стороны, высокая гибкость и универсальность ПЛИС позволяют стандартизовать такие устройства и выпускать их в виде унифицированных изделий, что позволит снизить стоимость в результате массового производства и упростить использование за счет разделения труда разработчиков.

В настоящей работе в качестве опытной платформы для исследования алгоритмов распознавания строк используются реконфигурируемые унифицированные вычислители (РУВ) [12]. Их применение позволяет в процессе проведения вычислительных экспериментов оперативно загружать разработанные структуры в ПЛИС, а также обеспечивать эффективное взаимодействие с центральным процессором компьютерной системы.

В работах [13-15] проанализированы предпосылки возникновения РУВ, обоснованы структура и состав таких устройств; исследованы возможные интерфейсы и типовые режимы обмена данными с центральным процессором вычислительной системы; проанализированы сложности использования и пути их решения; рассмотрены примеры реализации и организационные мероприятия, содействующие широкому распространению; исследованы категории сопутствующего программного обеспечения; обозначены перспективные области применения. Одной из областей применения, в которой преимущества РУВ способны проявиться в наибольшей степени, признаны задачи информационной безопасности.

Выводы

В настоящей работе проанализированы существующие программно-аппаратные системы распознавания строк в интенсивном потоке данных. Рассмотрены наиболее результативные подходы и методы аппаратного ускорения на базе ПЛИС, применяемые в системах обнаружения вторжений. Проанализированы их преимущества и недостатки, приведены ссылки на конкретные разработки.

Следует заметить, что наличие большого числа различных по своей природе направлений, конкурирующих в течение нескольких лет, которое не привело к выявлению лидирующего метода, существенно опережающего другие подходы по основным показателям, подводит к заключению, что техническую задачу распознавания строк в реальном масштабе времени в современных системах защиты информации пока еще не следует считать решенной.

Литература

1. A memory efficient multiple pattern matching architecture for network security / T. Song, W. Zhang, D. Wang, Y. Xue // IEEE INFOCOM. – 2008.
2. Лукацкий А.В. Обнаружение атак / Лукацкий А.В. – СПб. : БХВ-Петербург, 2001. – 624 с.
3. Jiang W. A FPGA-based Parallel Architecture for Scalable High-Speed Packet Classification / W. Jiang, V. Prasanna // Proc. of the International Conference on Application-Specific Systems, Architectures and Processors. – 2009. – P.24-31.
4. Палагин А.В. Реконфигурируемые вычислительные системы: Основы и приложения. / А.В. Палагин, В.Н. Опанасенко. – К. : Просвіта, 2006. – 280 с.
5. Flexible Software-Hardware Network Intrusion Detection System / R. Proudfoot, K. Kent, E. Aubanel, N. Chen. // Proc. 19th IEEE/IFIP International Symposium on Rapid System Prototyping. – 2008. – P. 182-188.
6. Sourdis I. Scalable multigigabit pattern matching for packet inspection / I. Sourdis, D. Pnevmatikatos, S. Vassiliadis. // Proc. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. – 2008. – v.16 n.2. – P. 156-166.
7. Jung H-J. Performance of FPGA Implementation of Bit-split Architecture for Intrusion Detection Systems / H-J. Jung, Z.K. Baker, V.K. Prasanna // Reconfigurable Architectures Workshop at IPDPS (RAW '06). – April 2006.
8. Lunteren J. High-Performance Pattern-Matching for Intrusion Detection / J. Lunteren // Proceedings of IEEE INFOCOM'06, (Barcelona, Spain). – April 2006.
9. Jiang W. Scalable Multi-Pipeline Architecture for High Performance Multi-Pattern String Matching / W. Jiang, V. Prasanna // IEEE International Parallel and Distributed Processing Symposium (IPDPS '10). – 2010.
10. Давиденко А.Н Алгоритмы распознавания строк в системах обнаружения вторжений на ПЛИС / А.Н Давиденко, С.Я. Гильгурт // Моделювання та інформаційні технології : зб. наук. пр. ІПМЕ НАН України. – Київ, 2010. – Вип. 58. – С. 103-109.
11. Rethinking Hardware Support for Network Analysis and Intrusion Prevention / V. Paxson, K. Asanovic, S. Dharmapurikar, J. Lockwood [etc.] // USENIX First Workshop on Hot Topics in Security (HotSec), (Vancouver, B.C.). – July 31, 2006.
12. Гильгурт С.Я. Решение задач распознавания с применением реконфигурируемых вычислителей / С.Я. Гильгурт, А.Н. Давиденко // Искусственный интеллект. – 2007. – № 1. – С. 123-131.
13. Гильгурт С.Я. Применение типовых устройств на базе программируемой логики для решения вычислительных задач / С.Я. Гильгурт // Параллельные вычисления и задачи управления : тез. докл. II международной конф., (4 – 6 окт. 2004 г.) – М. : Институт проблем управления им. В.А. Трапезникова РАН, 2004. – С. 514-530.
14. Гильгурт С.Я. Обзор современных реконфигурируемых унифицированных вычислителей / С.Я. Гильгурт // Моделювання та інформаційні технології : зб. наук. пр. ІПМЕ НАН України. – Вип. 49. – Київ: 2008. – С. 17-24.
15. Гильгурт С.Я. Анализ типовых режимов обмена данными с реконфигурируемыми вычислителями / С.Я. Гильгурт // Зб. наук. пр. ІПМЕ НАН України. – Київ, 2011. – Вип. 59. – С. 113-121.

Literatura

1. Song T., Zhang W., Wang D., Xue Y. A memory efficient multiple pattern matching architecture for network security. IEEE INFOCOM. 2008.
2. Lukackij A.V. Obnaruzhenieatak. SPb.: BHV-Peterburg. 2001. 624 s.
3. Jiang W. A. Proceedings of the International Conference on Application-Specific Systems, Architectures and Processors. 2009. P. 24-31.
4. Palagin A.V. Rekonfiguriruemyevychislitel’nyesistemy: Osnovyiprilozhenija. K.: “Prosvita”. 2006. 280 s.
5. Proudfoot R., Kent K., Aubanel E., Chen N..Proc. 19th IEEE/IFIP International Symposium on Rapid System Prototyping. 2008. P. 182-188.
6. Sourdis I. Proc. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2008. vol. 16 № 2. P. 156-166.
7. Jung H-J, Baker Z.K., Prasanna V.K. Reconfigurable Architectures Workshop at IPDPS (RAW’06). April 2006.
8. Lunteren J. Proceedings of IEEE INFOCOM’06. Barcelona. Spain. April 2006.
9. Jiang W. IEEE International Parallel and Distributed Processing Symposium (IPDPS’10). 2010.
10. Davidenko A.N, Gil’gurtS.Ja. Zb. nauk. pr. IPME NAN Ukrainy. Kyiv, 2010.Vyp. 58. S. 103-109.

11. Paxson V., Asanovic K., Dharmapurikar S., Lockwood J., etc. USENIX First Workshop on Hot Topics in Security (HotSec). Vancouver. B.C. July 31, 2006.
12. Gil'gurtS.Ja.,Davidenko A.N. Iskusstvennyj intellect. 2007. № 1. S. 123-131.
13. Gil'gurtS.Ja. Tez.dokl. II mezhdunarodnoj konf."Parallel'nye vychislenija i zadachi upravlenija". 4-6 okt. 2004 g. M.: Institut problem upravlenijaim. V.A. Trapeznikova RAN. 2004. S. 514-530.
14. Gil'gurt S.Ja. Modeljuvannja ta informacijnitehnologii. Zb. nauk. pr. IPME NAN Ukrainy. Vyp. 49. Kyiv. 2008. S. 17-24.
15. Gil'gurtS.Ja. Zb. nauk. pr. IPME NAN Ukrainy. Kyiv. 2011. Vyp. 59. S. 113-121.

S.Ya. Hilgurt

Hardware String Matching in the Intellectual Security Systems

Intellectual security systems are an effective tool for detecting various threats such as intrusion, virus and spam. The functions of such systems rely on multi-pattern string matching which scans the input stream to find all occurrences of a predefined set of string-based patterns rather than a single pattern. Due to the explosive growth of network traffic, multi-pattern string matching has been a major performance bottleneck in such systems which have to scan the incoming traffic in real time on fast links (e.g. 10 Gbps Ethernet and beyond).

General purpose CPUs are not able to take advantage of the available parallelism in the string matching tasks for information security. For these reasons, software-based intrusion detection systems (IDS) are often unable to keep up with the data rates of modern high-speed networks. As a means of improving the performance of such applications, researchers have turned to reconfigurable computing platforms where pattern-matching operations can be synthesized in custom hardware. Recently, many of IDS architectures have proposed for Field-Programmable Gate Arrays (FPGAs).

Following requirements must be met for the FPGA-based IDS solutions to be efficient and practical:

- High performance for a large pattern set;
- Good scalability;
- Deterministic throughput;
- Capability for dynamic update;
- Supporting all types of strings;
- Low memory requirement;
- Low implementation cost.

In this work, we propose FPGA-based Reconfigurable Unificated Accelerators (RUA) as a platform to solve and examine the problems mentioned above. Such devices combine advantages of software and hardware solutions. The application of RUA allows lowering essentially the cost of the technical solution due to their unification and mass production.

Many different techniques and their combinations can be realized and evaluated with the aid of RUA for the multi-pattern string matching:

- Deterministic / Nondeterministic Finite Automata;
- Discrete comparators;
- Content Addressable Memory;
- Variety of hash functions, in particular, Bloom filters.

Additionally, high-speed pattern matching is required for a wide variety of critical applications, including scanning through large data sets for data mining operations, low latency XML switching, DNA sequence matching, stateful packet inspection for QoS management etc.

Статья поступила в редакцию 24.10.2011.